DSR 환경에서의 다 모델 음성 인식시스템의 성능 향상 방법에 관한 연구

A Study on Performance Improvement Method for the Multi-Model Speech Recognition System in the DSR Environment

장현백*, 정용주**
Hyun-Baek Jang*, Yong-Joo Chung**

요 약

다 모델 음성인식기는 잡음환경에서 매우 우수한 성능을 보이는 것으로 평가되고 있다. 그러나 지금까지 다 모델 기반인식기의 성능시험에는 잡음에 대한 적응을 고려하지 않은 일반적인 전처리 방식이 주로 활용하였다. 본 논문에서는 보다 정확한 다 모델 기반인식기에 대한 성능 평가를 위해서 잡음에 대한 강인성이 충분히 고려된 전처리 방식을 채택하였다. 채택된 전처리 알고리듬은 ETSI (European Telecommunications Standards Institute)에서 DSR (Distributed Speech Recognition) 잡음환경을 위해서 제안된 AFE (Advanced Front-End) 방식이며 성능비교를 위해서 DSR 환경에서 좋은 성능을 나타낸 것으로 알려진 MTR (Multi-Style Training)을 사용하였다. 또한, 본 논문에서는 다 모델 기반인식기의 구조를 개선하여 인식성능의 향상을 이루고자 하였다. 기존의 방식과 달리 잡음음성과 가장 가까운 N개의 기준 HMM을 사용하여 기준 HMM의 선택시에 발생할 수 있는 오류 및 잡음신호의 변이에 대한 대비를 하도록 하였으며 각각의 기준 HMM을 훈련을 위해서 다수의 SNR 값을 이용함으로서 구축된 음향모델의 강인성을 높일 수 있도록 하였다. Aurora 2 데이터베이스에 대한 인식실험결과 개선된 다 모델기반인식기는 기존의 방식에 비해서 보다 향상된 인식성능을 보임을 알 수 있었다.

ABSTRACT

Although multi-model speech recognizer has been shown to be quite successful in noisy speech recognition, the results were based on general speech front-ends which do not take into account noise adaptation techniques. In this paper, for the accurate evaluation of the multi-model based speech recognizer, we adopted a quite noise-robust speech front-end, AFE, which was proposed by the ETSI for the noisy DSR environment. For the performance comparison, the MTR which is known to give good results in the DSR environment has been used. Also, we modified the structure of the multi-model based speech recognizer to improve the recognition performance. N reference HMMs which are most similar to the input noisy speech are used as the acoustic models for recognition to cope with the errors in the selection of the reference HMMs and the noise signal variability. In addition, multiple SNR levels are used to train each of the reference HMMs to improve the robustness of the acoustic models. From the experimental results on the Aurora 2 databases, we could see better recognition rates using the modified multi-model based speech recognizer compared with the previous method.

Keywords : Speech Recognition, Multi-model Speech Recognizer, Distributed Speech Recognition

I. 서 론

잡음 환경의 음성인식을 위한 많은 연구들이 지금까지 수행되어 왔으며, 이러한 노력들은 음성특징추출, 음성향상, 음성인식 모델 변환 등의 여러 가지 부분에서 이루어져 왔다[1][2][3]. 이들은 각자 독립적으로 사용되거나 결합적으로 사용되어서 잡음환경에서 인식성능의 향상을 이루는데 크게 기여하고 있다. 이와 같은 기존의 잡음 환경의 음성인식 방식들과는 다른 접근 방법으로서, 최근에 들어서는 하나의 음향모델을 사용하는 대신에 실제 상황에서 발생 가능한 잡음의 종류를 미리 예측하여 잡음의 종류별로 음향모델들을 미리 훈련하고 이들을 인식시에 함께 이용하는 다 모델 기반의 음성인식시스템이 제안되어 매우 효과적인 결과를 얻을 수 있음이 알려져 있다[4].

음성인식시스템이 동작하는 실제 상황은 매우 다양한 잡음 환경을 포함할 수 있으며 그 중에서도 DSR

* 희성전자(주) ** 계명대학교 전자공학과
투고 일자 : 2010. 3. 2 수정완료일자 : 2010. 4. 28
게재확정일자 : 2010. 4. 29
* 이 논문은 2009년도 정부의 재원으로 한국과학재단의 지원을 받아 수행된 연구임(No. 2009-0075480)

(distributed speech recognition)은 가장 대표적이라 할 수 있다. ETSI (European Telecommunications Standards Institute)에서는 DSR 잡음환경에 대비한 최적의 음성 전처리 (front-ends) 방식을 제안하였으며 이 방식은 잡음환경 음성인식을 위한 가장 대표적인 전처리 알고리듬으로 간주되어 Aurora 데이터베이스를 이용한 잡음 환경 음성인식시스템의 성능 비교시에 기본적인 전처리 방식으로 이용되고 있다[5][6].

ETSI에서 제안한 DSR을 위한 전처리 방안은 크게 2 가지로 나눌 수 있다. 그 첫 번째 방식(FE)은 일반적인 MFCC (mel frequency cepstral coefficients) 기반의 특징 추출을 사용하고 있다[5]. FE 방식은 잡음환경에서 그리 좋은 성능을 보이지 못하였고 따라서 보다 잡음환경에 적합하도록 설계된 두 번째 방식(AFE: Advanced Front End)이 제안되었다[6]. AFE 방식은 MFCC 특징을 추출하는 것 외에도 VAD(voice activity detection)와 Wiener 필터링을 응용한 잡음제거 알고리듬을 포함함으로서 잡음신호에 잘 적응하도록 하였으며 잡음제거 이후에는 SNR(signal to noise ratio) 값이 높은 부분은 강조하고 낮은 SNR값의 영역은 약화시킴으로서 전체적으로 SNR값이 향상되도록 하는 과정을 포함하고 있다. 이와 같이 단순한 MFCC의 추출 외에도 잡음제거와 SNR값 향상을 통하여 AFE는 FE에 비해서 연속숫자음 인식에 있어서 53%의 인식오류 감소의 효과가 있음이 알려져 있다[7].

다 모델 기반의 음성인식기는 기존의 잡음음성인식 방식에 비해서 보다 나은 성능을 보였는데, PMC 나 JA 방식 등의 인식모델 변환 방식이나 SS(spectral subtraction) 과 같은 음성향상 방법 등에 비해서도 우수한 성능을 보임을 알 수 있었다. 특히, 잡음 환경에서 매우 우수한 성능을 보이고 있는 MTR(Multi-style TRaining) 훈련 방식에 비해서도 더 나은 성능을 보이는 것으로 알려져 있다[4].

그러나 현재까지의 다 모델 기반 음성인식기의 성능 평가시에는 DSR 전처리 방식 중에서 FE 방식을 이용하였고 AFE를 사용한 경우의 다 모델 기반 인식기의 성능에 대한 평가는 이루어지지 않았다. 따라서 본 연구에서는 잡음환경을 위한 전처리 방식 중에서 가장 대표적이며 성능이 매우 우수한 AFE를 사용한 경우의 다 모델 기반 인식기의 성능을 비교 검토 할 것이다. 또한 다 모델 기반 인식기의 성능 향상을 위한 몇몇 방안을 제시하고자 한다.

다 모델 기반의 음성 인식기에서는 인식음성에 포함된 잡음신호가 훈련된 잡음모델들 중에서 어느 것과 가장 유사하는지를 찾아야 하는데 이 과정에서는 항상 오류가 발생하고 또한 찾아진 잡음모델은 인식 잡음신호와 어느 정도 차이가 나게 마련이므로 결과적으로 인식성능의 저하를 가져오게 된다. 따라서 이러한 문제를 해결하기 위해서 본 연구에서는 하나의 잡음음성모델을 구성하기 위한 SNR범위를 확대시키고 가장 가까운 하나의 잡음음성모델을 선택하는 대신 다수의 모델을 활용하는 방안을 이용하여 다 모델 인식기의 강인성을 향상시키고자 한다.

II. 다 모델기반의 음성인식기

2.1 개선된 다 모델인식기

다 모델 기반의 음성인식기에서는 인식시에 주변 환경에서 발생할 것으로 예상되는 잡음종류별로 그리고 다양한 SNR값에 대하여 각각의 기준HMM 모델을 훈련과정 중에 구성하도록 하며, 인식시에는 이러한 기준HMM 모델들 중에서 인식 잡음음성신호에 가장 적합한 것을 선택하여 최종 인식모델로 삼는다. 이러한 방식은 하나의 기준HMM을 고려하는 기존의 방식에 비해서 보다 다양한 잡음환경에 적응할 수 있는 강인성을 향상 시킬 수 있다는 이점이 있다.

기존의 다 모델 기반의 음성인식기에서는 인식 잡음음성에 가장 적합한 기준HMM 모델을 찾기 위해서는 인식 잡음음성의 신호 대 잡음비(SNR)를 추정해야 할 뿐만 아니라 인식 잡음음성에 포함된 인식 잡음신호에 대한 분류를 하는 것이 필요하였다[4]. 본 연구에서는 기존의 다 모델 기반 인식기를 다소 개선하였으며 전체 블록다이어그램은 <그림 1>에 나타나 있다.

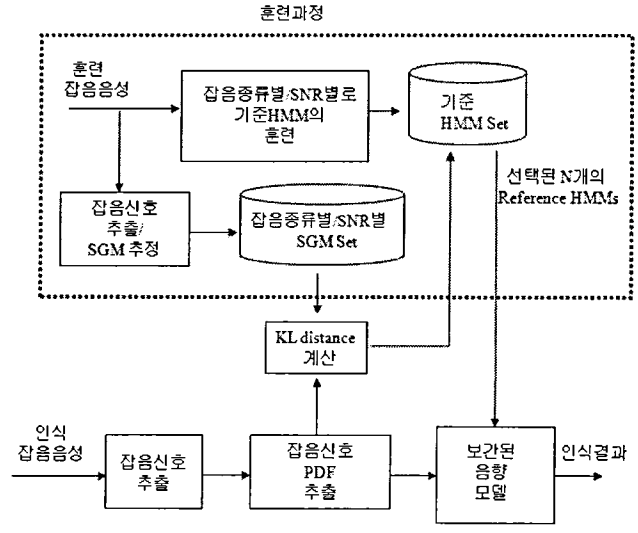

그림 1. 개선된 다 모델 기반 인식기의 구조
Fig. 1. The Architecture of the Improved Multi-model based Speech Recognizer

우선, 인식 잡음음성으로부터 추출된 인식 잡음신호와 기준HMM 모델과의 유사성을 측정하기 위해서 Kullback-Leibler (KL) distance 값을 사용하였다. 가장 유사도가 높은 하나의 기준 HMM만을 이용하던 기존의 방식에 비해서 개선된 방식에서는 유사도가 큰 N개의 기준 HMM을 상호보간(interpolation)함으로서 인식성능의 향상을 꾀하였다. 이러한 상호보간은 유사도 추정에 있어서 자주 발생하는 오류에 대한 대비를 위한 방법일 뿐만 아니라 하나의 기준 HMM을 사용하는 것 보다는 유사한 여러 개의 기준 HMM을 함께 사용함으로서 잡음신호의 변이에 대한 인식기의 강인성을 높일 수 있도록 하기 위함이다. N개

의 기준HMM 각각의 확률밀도 함수가 $f_i(O), i=1,\cdots,N$ 이라고 하면 상호보간이 이루어진 후의 확률밀도 함수 $f(O)$는 다음과 같다.

$$f(O)=\sum_{i=1}^{N}\alpha_i f_i(O) \quad (1)$$

식(1)에서 O는 잡음음성 특징벡터를 의미하며 α_i 값은 각각의 확률밀도 함수에 대한 가중치 값이다. 본 연구에서는 $\alpha_i=\frac{1}{N}$로 두어서 모든 기준HMM의 비중을 동일하게 두었다. 실제 인식실험에서는 가중치 α_i 값에 차등을 둔 경우와 동일하게 둔 경우를 비교한 경우 인식성능에 큰 차이가 없음을 알 수 있었다. 기존의 연구에서도 Xu 등은 기준HMM 모델과 MTR 모델의 보간을 위해서 실험적으로 가중치 α_i를 정한바 있다[4].

개선된 인식기에서는 인식 잡음음성신호로부터 추출된 잡음신호에 가장 근접한 N개의 기준HMM을 선택하기 위하여 훈련과정에서 미리 가정된 잡음신호의 종류 및 SNR 레벨 별로 단일모드 가우시안 모델 (SGM: Single mode Gaussian Model)에 대한 확률밀도 함수를 추정하였다.

D-차원의 잡음신호 특징벡터 n 대해서 평균벡터 μ와 공분산 행렬 Σ를 갖는 단일모드 가우시안 확률밀도 함수는 다음과 같다.

$$p(n)=\frac{1}{(2\pi)^{D/2}|\Sigma|^{D/2}}exp\{-\frac{1}{2}(n-\mu)'(\Sigma)^{-1}(n-\mu)\} \quad (2)$$

각각의 잡음종류 및 SNR 레벨별로 잡음신호의 특징벡터가 얻어지면 각 확률밀도함수의 평균벡터 μ 및 공분산 행렬 Σ 는 expectation-maximization (EM) 방식에 의해서 주어진 잡음신호에 대해서 최대우도를 갖도록 구해지게 된다. 훈련과정에서 SGM이 추정되고 나면 인식 잡음음성신호로부터 추출된 잡음신호의 확률밀도 함수와 SGM간의 KL distance를 구하여 그 값이 가장 작은 N개의 SGM을 선정함으로서 해당하는 기준 HMM을 선택할 수 있게 된다. 두 개의 단일 모드 가운시안 확률밀도 함수 $N_1(\mu_1,\Sigma_1), N_2(\mu_2,\Sigma_2)$ 간의 KL distance 값은 다음식과 같다[8].

$$KLD(N_1,N_2)=\frac{1}{2}\sum_{i=1}^{D}\left[\log\left(\frac{(\Sigma_1)_{ii}}{(\Sigma_2)_{ii}}\right)+\frac{((\mu_2)_i-(\mu_1)_i)^2}{(\Sigma_1)_{ii}}+\left(\frac{(\Sigma_2)_{ii}}{(\Sigma_1)_{ii}}-1\right)\right] \quad (3)$$

식(1)과 식(3)에 나타난 바와 같이 기준HMM을 상호보간하는 제안된 방식은 확률밀도함수 $f(O)$를 구하는 과정에서 기존의 방식보다 많은 계산량을 요구하며 KL distance를 구해야 하는 복잡성을 야기하는 단점이 있다. 이러한 문제점은 향후 연구를 통해서 개선되어야 할 것으로 생각된다.

앞에서 언급한 것처럼 훈련과정에서는 다양한 잡음종류별 및 SNR 레벨별로 별도의 기준 HMM을 구성하게 된다. 이때, 같은 종류의 잡음에 대해서도 다양한 SNR 레벨값에 따라서 각기 다른 기준HMM을 구성하게 되는데 기존의 연구에서는 보다 분별적인 음향모델을 구축한다는 의미에서 SNR 레벨 값의 간격을 가급적 조밀하게 설정하였으나 이는 유사도 추정오류가 발생하는 경우 인식성능이 다소 민감하게 반응하는 문제점을 낳았다[8]. 따라서 개선된 인식기에서는 SNR 레벨 값의 범위를 기존보다 다소 넓게 잡아서 기준 HMM을 구성함으로서 인식기의 성능향상을 이루었다. 이에 대한 보다 자세한 사항은 4장의 인식실험 결과에서 다루고자 한다.

2.2 DSR 표준 전처리

DSR에서의 음성인식을 위하여 ETSI에서는 2가지 종류의 표준 전처리 과정을 제안하였다. DSR을 위한 첫 번째 표준 전처리 과정인 ES 201 108은 2000년에 발표되었다. 이 방식은 FE라 불리어 지며 MFCC기반의 음성특징을 생성하는 특징추출 부분과 음성데이터의 채널 전송을 위한 인코딩 부분으로 크게 나누어진다. <그림 2>에는 FE 방식의 전처리 과정에 대한 블록다이어그램이 나타나 있다. 본 연구에서는 잡음음성에 대한 영향만을 고려하기 위하여 채널 코딩 부분은 제외하고 <그림 2>의 특징추출 부분만을 구현하여 실험하였다.

FE에서의 MFCC 특징 추출과정에는 음성신호의 dc offset 보상, 고주파 성분을 위한 프리엠퍼시스 (pre-emphasis), 스펙트럼 크기 계산, 멜스케일(mel-scale)의 필터뱅크 출력의 로그 값 추출 그리고 마지막으로 DCT (discrete cosine transformation) 변환을 통한 MFCC 계산 과정이 포함되어 있다. 특징추출은 매 프레임 별로 13차의 켑스트럼 벡터와 1차의 log energy 값을 생성하여 전체적으로 14차의 특징벡터를 만들어 낸다.

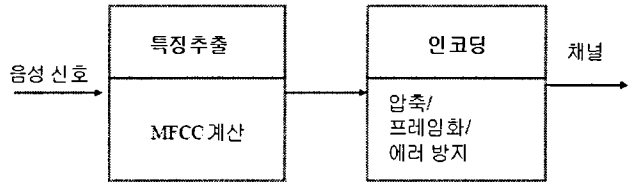

그림 2. FE 전처리 방식의 개요도
Fig. 2. Block diagram of the FE front-end

앞에서 소개된 FE 전처리과정은 지하철이나 자동차 환경의 인식에서 성능이 좋지 못한 것으로 알려져 있다. 따라서 FE를 개선하여 잡음에 강인한 전처리 과정인 AFE가 2002년에 제안되었으며 ESTI 표준 문서인 ES 202 050에 소개되었다.

<그림 3>에는 AFE 전처리 과정에 대한 블록다이어그램이 나타나 있다. 예상했듯이 잡음 환경에서의 인식성능 향상을 위하여 FE에 비하여 몇 가지 모듈이 추가되어 있음을 <그림 3>을 보면 알 수 있다. Wiener 필터링 기반의 잡음제거, SNR을 향상시키기 위한 waveform 처리 그리고 컨벌루션 잡음을 보상하기 위한 blind equalization 및 VAD (voice activity detection) 등이 새로이 추가 되었다.

MFCC계산 모듈은 기존의 FE와 동일하게 구성된다.

FE 전처리를 이용한 기존의 연구에서 다 모델 인식기의 성능이 타 방식에 비해 우수함이 알려져 있다[4]. 그러나 AFE 전처리를 사용한 경우에도 다 모델 인식기의 성능의 우수성이 유지되는지를 비교 검토할 필요가 있다고 생각된다. 따라서 본 연구에서는 앞에서 설명된 개선된 다 모델 인식기를 이용하여 AFE 전처리를 사용한 경우의 인식성능을 타 방식과 비교하고 다 모델 인식기의 성능향상 방법에 대하여 논의하고자 한다.

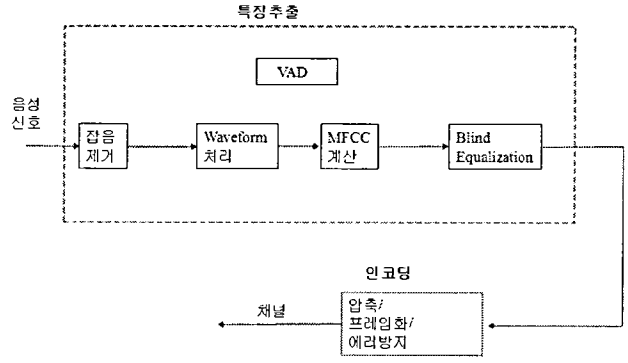

그림 3. AFE 전처리 방식의 개요도
Fig. 3. Block diagram of the AFE front-end

III. 음성 데이터 및 인식 시스템

본 연구를 위하여 잡음음성인식을 위한 Aurora 2 데이터베이스를 사용하였다. Aurora 2 데이터베이스는 TI digit 데이터베이스에 인공적으로 부가잡음을 더해주고 채널왜곡을 인가한 것이다. 훈련은 CLEAN 과 MTR 방식이 있으며 전자는 HMM의 훈련시에 깨끗한(clean) 음성데이터 만을 이용하고 후자는 깨끗한 음성과 함께 여러 가지 종류의 잡음음성을 다양한 SNR 값으로 구성한 후 이를 모두 이용하여 HMM을 훈련하는 방식이다. 인식을 위해서는 3가지 종류의 음성데이터 Set 이 사용된다. Set A는 인식잡음이 훈련시 사용된 잡음의 종류와 같은 경우(Subway, Babble, Car, Exhibition noises)이며 Set B는 인식잡음이 훈련 잡음과 다른 경우(Restaurant, Street, Airport, Train-Station noises) 이고 Set C는 앞의 부가 잡음 외에도 채널 왜곡이 인위적으로 조성된 경우이다.

음성특징을 위해서는 ETSI에서 DSR 환경에서의 음성인식을 위하여 제안된 AFE를 사용하였고 0차의 cepstral 계수를 제외한 12차의 MFCC(Mel-frequency Cepstral Coefficient)에 1차의 로그에너지를 포함한 13차의 특징벡터를 기본으로 하고 여기에 delta 와 acceleration 계수를 추가하여 전체 39차의 특징벡터를 사용하였다[8].

숫자에 대한 HMM은 3개의 Gaussian 확률밀도함수 성분을 가지는 16개의 상태들로 이루어지며 묵음에 관한 HMM 모델은 6개의 Gaussian 성분을 가지는 3개의 상태로 이루어진다. 한편 1개의 상태를 가지는 짧은 묵음에 관한 HMM 모델도 구성되며 이는 묵음모델의 가운데 상태와 동일하게 모델링 된다. 인식기는 본 연구실에서 자체적으로 개발하였으며 Baum-Welch 기반의 훈련과정을 통해서 HMM 모델을 구성하고 Viterbi 알고리듬을 이용하여 인식결과를 얻도록 하였다[9][10].

IV. 인식 실험 및 고찰

먼저 AFE와 FE 전처리간의 잡음환경에서의 인식성능을 비교 검토하기 위해서 표 1에서는 FE와 AFE를 사용한 경우에 Aurora 2 데이터베이스에 대한 인식 결과를 보여 주고 있다. CLEAN 방식과 MTR 방식을 각각 사용하였다.

표 1. FE 와 AFE 전처리 방식의 성능 비교 (WER(%)).
Table 1. Performance comparison between FE and AFE front-ends

훈련방식 \ 전처리방식		FE	AFE
CLEAN	set A	37.43	13.67
	set B	42.94	14.58
	set C	33.08	15.36
	평균	**38.75**	**14.37**
MTR	set A	12.55	8.51
	set B	13.71	8.94
	set C	17.03	9.83
	평균	**13.91**	**8.95**

<표 1>에서 보면 CLEAN 훈련방식을 채택한 경우 FE의 단어오인식율(WER: word error rate)는 setA, setB, setC에 대한 결과를 평균한 결과 38.75(%) 인 반면에 AFE의 경우에는 14.37(%)임을 알 수 있으며 이는 약 63(%)의 오인식율 감소를 의미한다. 잡음에 강인한 훈련방식인 MTR의 경우에도 AFE는 FE에 비하여 약 35(%)의 오인식율 감소효과를 가져 옴을 알 수 있다. <표 1>의 Aurora 2 데이터베이스에 대한 인식실험결과에서 AFE는 CLEAN 과 MTR 훈련방식 모두에서 FE 보다 월등히 우수한 성능을 나타냄을 알 수 있다. 따라서 기존의 연구에서 FE 전처리를 사용한 경우에 다 모델 기반 인식기가 잡음 환경에서 우수한 성능을 보임을 알 수 있었지만 이에 대한 보다 정확한 평가를 내리기 위해서는 AFE를 사용한 경우에 성능비교를 해야 할 것이다.

<표 2>에는 다 모델 기반인식기의 성능을 AFE 전처리를 사용한 경우에 보간되는 기준HMM의 개수 N=1, 2, 4, 6 인 경우 각각에 대해서 나타내었다. <표 2>에서는 기존의 다 모델기반인식기에 비해서 (N=1 인 경우) 보간되는 기준 HMM의 개수를 증가시키는 경우 인식성능이 다소 향상됨을 알 수 있다. N=4인 경우에 10.71(%)의 WER를 나타내어 가장 나은 보임을 알 수 있다. N이 증가하는 경우에 set A, B, C 모든 경우에서 전반적으로 인식성능이 향상되지만 특히 set C의 경우에서 많은 성능향상이 이루어짐을 알 수 있다. 예를 들어, N=6인 경우에 보면 set A와 set B에서는 N=4인 경우에 비해서 성능의 저하가 다소 발생하

지만 set C의 경우에는 여전히 인식성능이 향상됨을 알 수 있다. <표 2>에서와 같이 N이 어느 정도 증가하면 인식성능이 향상되는 것은 KL distance를 통해서 가장 유사한 기준HMM을 찾는 과정에 오류가 생길 개연성이 많고 또한 잡음신호의 변이에 대하여 단일 기준HMM에 비해서 상호보간된 HMM이 보다 적절히 대응할 수 있기 때문이다.

표 2. 보간되는 기준HMM의 개수의 변화에 따른 다 모델 기반 인식기의 성능비교(WER(%)).
Table 2. Performance comparison of the multi-model based recognizer as the number of interpolated rereference HMMs changes.

보간되는 기준HMM의 수 (N)	set A	set B	set C	평균
1	9.28	13.24	9.95	11.00
2	9.16	13.21	9.49	10.85
4	9.17	13.15	8.92	10.71
6	9.18	13.32	8.8	10.76

개선된 다 모델 인식기에서는 <표 2>에서 보인바와 같이 기준HMM의 보간을 통해서 인식성능을 향상시킬 뿐 아니라 각각의 기준HMM에 해당하는 SNR의 범위를 기존에 비해서 확대해줌으로서 인식성능의 향상을 꾀하였다. <표 3>에는 기존의 방식과 비교해서 SNR의 범위를 확대한 두 가지 경우인 SNRMERG와 SNRMERG2에서의 SNR의 값의 범위를 나타내었다.

기존의 방식에서는 SNR값이 0dB, 5dB, 10dB, 15dB, 20dB, 25dB, 30dB 인 경우 각각에 대해서 독립적으로 기준HMM을 구성하였다. 반면에 개선된 방식인 SNRMERG에서는 0dB와 5dB, 10dB와 15dB, 20dB와 25dB를 각각 합쳐서 하나의 기준HMM을 구성함으로서 잡음종류별 기준HMM의 수를 7개에서 4개로 줄였다. 또한 SNRMERG2에서는 SNRMERG와 유사하게 이웃하는 SNR값을 묶어서 하나의 기준HMM을 구성하되 기준HMM 서로 간에 SNR 값이 겹치도록 하였다.

표 3. 기준HMM의 훈련시 적용된 SNR 값들
Table 3. SNRs applied in training the referece HMM

	기존방식	SNRMERG	SNRMERG2
기준 HMM별 SNR값들	{0},{5}, {10},{15}, {20},{30}	{0,5}, {10,15}, {20,25},{30}	{0,5},{5,10}, {10,15}, {15,20}, {20,25},{25,30}, {30}
잡음종류별 기준 HMM의 개수	7	4	7

<표 4>에는 SNRMERG 와 SNRMERG2 방식을 사용한 경우의 다 모델 인식기의 성능을 비교하여 나타내었다.

표 4. SNRMERG와 SNRMERG2 방식을 사용한 경우의 다 모델 인식기의 성능비교 (WER(%)).
Table 4. Performance comparison of the SNRMERG and SNRMERG2 in the multi-model based recognizer.

	N	set A	set B	set C	평균
SNRMERG	1	9.01	13.12	9.75	10.80
	2	8.60	13.04	9.02	10.46
	4	8.94	13.01	8.49	10.48
	6	9.17	13.07	8.49	10.59
SNRMERG2	1	8.80	12.72	9.66	10.54
	2	8.63	13.02	9.38	10.54
	4	8.70	13.17	9.10	10.57
	6	8.93	13.28	8.66	10.62

<표 4>에서 알 수 있듯이 SNRMERG와 SNRMERG2를 사용한 경우에 기존의 방식에 비해서 다소 나은 성능을 보임을 알 수 있다. <표 2>의 기존의 방식에서는 N=1인 경우에 11(%)의 단어오인식율을 나타내고 있으나 <표 4>의 SNRMERG와 SNRMERG2의 경우에는 N=1인 경우 각각 10.8(%)와 10.54(%)를 나타내어 인식율의 향상이 나타남을 알 수 있다. 물론 <표 2>의 기존의 방식에서도 보간되는 기준 HMM의 개수를 증가시킴으로서 N=4인 경우에 오인식율이 10.71(%)까지 향상이 일어나는 것을 볼 수 있는데 이를 통해서 기준HMM의 보간이 기준HMM의 훈련과정에서 SNR 범위를 확대하는 것과 마찬가지로 성능향상의 효과를 나타내는 것을 확인 할 수 있다. 또한 SNRMERG의 경우에도 기준HMM의 보간을 통해서 더욱 성능이 향상됨을 확인할 수 있다. SNRMERG는 N=2의 경우에 10.46(%)의 오인식율을 나타내며 SNRMERG2의 경우에는 10.54(%)의 오인식율을 나타내어 두 가지 방식이 나타내는 최고의 인식율에는 큰 차이가 없다고 할 수 있으나, SNRMERG2의 경우에는 N=1에서 최고의 인식율을 나타내므로 보간과정이 따로 필요없다는 장점이 있다.

개선된 다 모델 인식기의 성능을 MTR 훈련방식과 비교하였으며 그 결과가 <표 5>에 나타나 있다

SNRMERG(N=2)와 SNRMERG2(N=1)는 개선된 다 모델 기반 인식기에서 가장 성능이 우수한 경우들인데 이를 MTR과 비교해보면 MTR이 개선된 다 모델 기반 인식기들에 비해서 우수한 성능을 보임을 알 수 있다. 기존의 FE 전처리 인식기를 사용한 경우에는 다 모델 기반 인식기의 성능이 MTR에 비해서 우수했던 것과 정반대의 결과를 보이고 있음을 알 수 있는데, 이는 AFE 전처리 방식에서는 잡음제거 알고리듬과 SNR 향상 알고리듬이 포함되어 있어서 다 모델 기반 인식기가 잡음에 강인한 효과가 상대적으로 FE 전처리를 사용할 경우에 비해서 약해지기 때문인 것으로 생각된다. <표 5>에서는 SNRMERG2(N=1)과 MTR를 상호보간한 결과를 나타내고 있는데, 이는 다 모델 기반 인식기와 MTR 방식의 결합을 통해서 서로간의 보완이 가능할 것으로 판단되기 때문이다. <표 5>의 결과를 보면 이러한 상호보간 방식이 전체적으로는 MTR보다 저조하지만

set B를 제외한 set A 와 set C에서는 오히려 보간방식이 MTR 보다 향상된 성능을 보임을 알 수 있다. Set B에서 보간 방식의 성능이 MTR에 비해서 상당히 저조한데 이는 다 모델 인식기의 기준HMM의 훈련에 포함되지 않은 잡음 신호가 set B에 있어서 이에 크게 영향을 받은 때문이다. 따라서 set B에 대한 성능보완이 다소 이루어지면 개선된 다 모델 기반 인식기와 MTR 방식을 상호보간 함으로서 최적의 인식성능을 나타낼 수 있을 것이다.

표 5. 개선된 다 모델 인식기와 MTR 방식의 성능비교 (WER(%)).

Table 5. Performance comparison between the improved multi-model based recognizer and the MTR method

	set A	set B	set C	평균
기존의 다 모델 인식기	9.28	13.24	9.95	11.00
SNRMERG(N=2)	8.94	13.01	8.49	10.48
SNRMERG2(N=1)	8.80	12.72	9.66	10.54
MTR	8.51	8.94	9.83	8.95
SNRMERG2(N=1)+ MTR	8.21	10.66	8.46	9.24

V. 결 론

본 연구에서는 다 모델 기반인식기의 구조를 개선하여 인식잡음신호에 가장 가까운 1개의 기준HMM을 선택하는 대신 KL distance 기반의 N개 기준HMM을 선택할 수 있도록 하였으며 또한 기준HMM의 훈련시에 SNR의 범위를 확대함으로서 기준HMM의 강인성을 높일 수 있도록 하였다. 이러한 개선된 방식을 통해서 기존의 다 모델 기반인식기의 성능이 향상됨을 확인할 수 있었다. 또한 개선된 다 모델 기반 인식기를 기존의 MTR 방식과 상호보간 함으로써 더욱 강인한 다 모델기반 인식기를 구축할 수 있었으며 이를 통해서 다 모델기반 인식기의 효용성을 입증하였다. 그러나 AFE 특징을 사용한 경우에 기존의 MTR 방식과 비교한 결과 다소 저조한 성능을 보였는데 이를 보완하기 위해서는 훈련중에 포함되지 않은 잡음신호에 대한 적응능력을 향상시키는 방법에 대한 연구가 향후 필요하리라 생각된다.

참 고 문 헌

[1] Gales, M. J. F., Model Based Techniques for Noise-Robust Speech Recognition, Ph.D. Dissertation, University of Cambridge. 1995.

[2] Moreno, P. J., Speech Recognition in Noisy Environments, Ph.D. Dissertation, Carnegie Mellon University, 1996.

[3] Ball, S. F., "Suppression of Acoustic Noise in Speech Using spectral subtraction", IEEE Trans. Acoust., Speech, Signal Process., vol.27, pp.113-120, 1979.

[4] Xu, H., Tan, Z.-H., Dalsgaard, P., Lindberg, B., "Robust Speech Recognition on Noise and SNR Classification - a Multiple-Model Framework", Proc. Interspeech, 2005.

[5] ETSI Draft Standard Doc. Speech Processing, Transmission and Quality Aspects (STQ); Distributed Speech Recognition; Front-End Feature Extraction Algorithm; Compression Algorithm, ETSI Standard ES 202 108, 2000.

[6] ETSI Draft Standard Doc. Speech Processing, Transmission and Quality Aspects (STQ); Distributed Speech Recognition; Advanced Front-End Feature Extraction Algorithm; Compression Algorithm, ETSI Standard ES 202 050, 2002.

[7] Macho, D., Mauuary, L., Noe, B., Cheng, Y., Eahey, D., Jouvet, D., Kelleher, H., Pearce, D., Saadoun, F., "Evaluation of a Noise-Robust DSR Front-End on Aurora Databases", Proc. ICSLP, pp.17-20, 2002.

[8] Juang, B. H. and Rabiner, L. R., "A Probabilistic Distance Measure for Hidden Markov Models", AT&T Technology Journal, pp. 391-408, 1984.

[9] 정용주, "연속 잡음 음성 인식을 위한 다 모델 기반 인식기의 성능 향상에 대한 연구", 음성과학, 제15권 제2호, pp. 55-65, 2008.

[10] 김희근, 정용주, "AURORA DB를 이용한 잡음 음성 인식실험을 위한 Segmental K-means 훈련방식의 기반인식기의 구현", 말소리, 제57호, pp 113-122, 2006.

장 현 백 (Hyun-baek Jang)

2010년 2월 계명대학교 전자공학과 (공학사)
2010년 3월~현재 희성전자(주)
관심분야 : 신호처리, 음성인식

정 용 주 (Yong-joo Chung)

1988년 2월 서울대학교 전자공학과 (공학사)
1990년 2월 KAIST 전기및전자공학과 (공학석사)
1995년 8월 KAIST 전기및전자공학과 (공학박사)
1995년 9월 ~ 1999년 2월 LG전자 선임연구원
1999년 3월 ~ 현재 계명대학교 전자공학과 부교수
관심분야 : 음성인식, 신호처리